

TIME: A Multi-level Benchmark for Temporal Reasoning of LLMs in Real-World Scenarios

Shaohang Wei, Wei Li, Feifan Song, Wen Luo, Tianyi
Zhuang, Haochen Tan, Zhijiang Guo, Houfeng Wang

Peking University Huawei Noah's Ark Lab

NeurIPS 2025 Track on Dataset and Benchmarks (Spotlight)

- Effective temporal reasoning is crucial for Large Language Models (LLMs) to process and comprehend complex events with human-like understanding, particularly in applications requiring integration of historical data and real-time progress tracking. LLMs still face challenges in managing temporal understanding in reality.
- 3 key complex challenges for temporal reasoning in real-world context.
 - Intensive temporal knowledge information
 - Rapid evolution of event details over time
 - Complexity of temporal dependencies in social interactions
- However, existed benchmarks, like *TimeBench* and *TRAM* primarily focus on simplified scenarios, such as basic temporal commonsense and relationships within short texts and simple QA tasks.

- Temporal reasoning constitutes a **hierarchical** framework of **fine-grained** abilities, which is also different from other reasoning tasks that focus on singular capabilities, but is still ignored by current works.
 - *TReMu* involves only neuro-symbolic temporal reasoning while neglecting temporal computation.
 - *TCELongBench* overlooks fundamental temporal concept understanding.
- In contrast, a robust evaluation framework should encompass both basic temporal abilities and complex event-event temporal reasoning, necessitating the development of a new comprehensive benchmark.
- We propose TIME, a multi-level comprehensive benchmark for evaluating temporal reasoning in LLMs across diverse real-world scenarios, comprising 38,522 instances. And we select a subset TIME-Lite for future works.

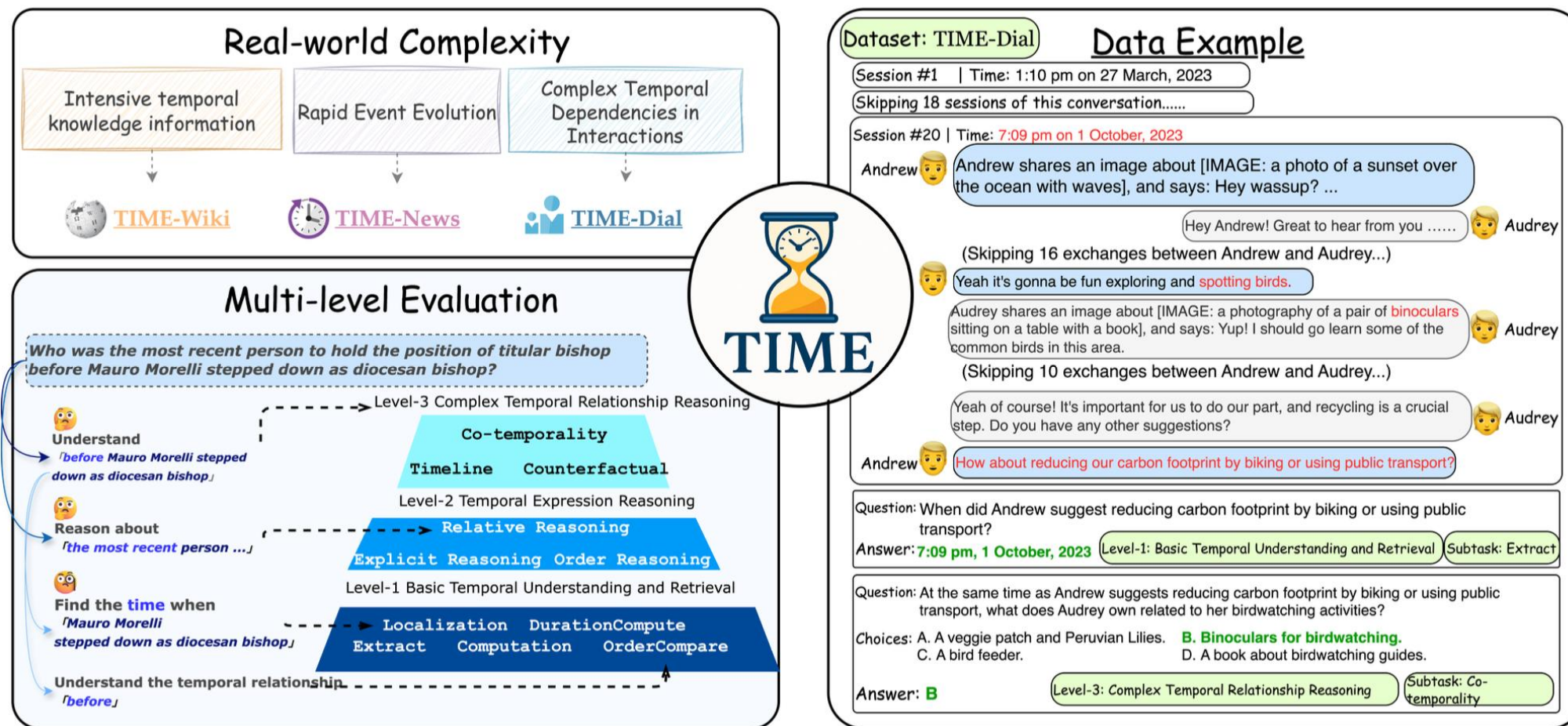


Figure 1: An overview of TIME. The top-left block illustrates three key challenges of real-world complexity and their corresponding dataset construction. The bottom-left quadrant depicts a three-level tasks. One data example from TIME-DIAL is shown on the right.

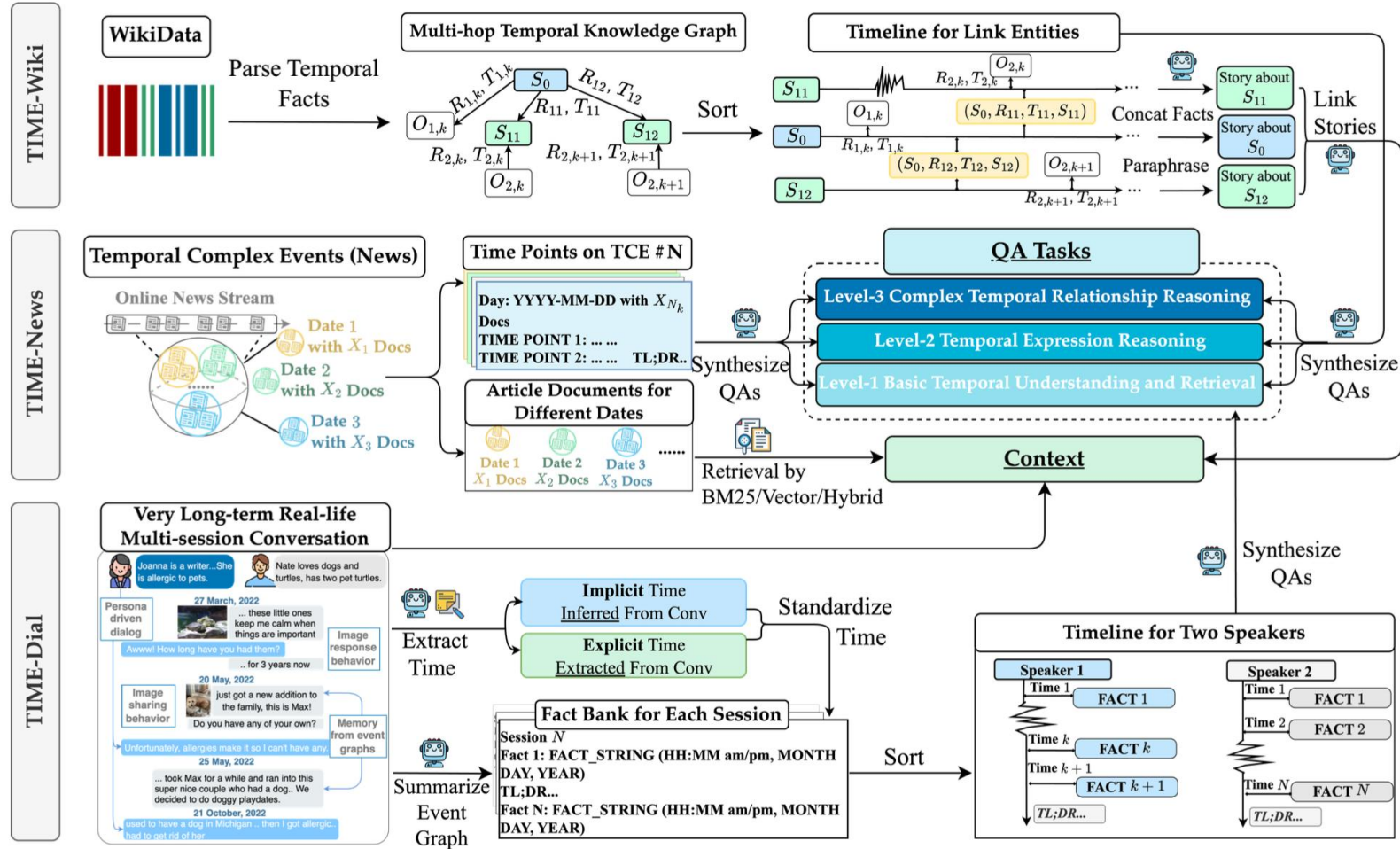


Figure 2: Dataset construction pipeline for TIME. In the process of QA synthesis for each sub-dataset, we first collect temporal facts (temporal knowledge graphs for TIME-WIKI, time points for TIME-NEWS, fact bank for TIME-DIAL). Then timelines are generated for QA data synthesis.

Table 10: Dataset statistics. The table displays the number of instances for each dataset and task category. Task abbreviations are: Ext. (Extract), Loc. (Localization), Comp. (Computation), D.C. (Duration Compare), O.C. (Order Compare); E.R. (Explicit Reasoning), O.R. (Order Reasoning), R.R. (Relative Reasoning); C.T. (Co-temporality), T.L. (Timeline), C.F. (Counterfactual).

Dataset	All Tasks	Level 1					Level 2			Level 3		
		Ext.	Loc.	Comp.	D.C.	O.C.	E.R.	O.R.	R.R.	C.T.	T.L.	C.F.
TIME	38522	1480	3546	3376	3401	3549	3537	3538	3537	3513	5508	3537
TIME-WIKI	13848	1261	1299	1126	1151	1299	1287	1288	1287	1263	1300	1287
TIME-NEWS	19958	0	1800	1800	1800	1800	1800	1800	1800	1800	3758	1800
TIME-DIAL	4716	219	447	450	450	450	450	450	450	450	450	450
TIME-LITE	943	60	90	78	86	90	90	90	90	90	89	90
TIME-LITE-WIKI	322	30	30	24	28	30	30	30	30	30	30	30
TIME-LITE-NEWS	299	0	30	30	30	30	30	30	30	30	29	30
TIME-LITE-DIAL	322	30	30	24	28	30	30	30	30	30	30	30

- 24 Models
 - Vanilla models (non-reasoning models)
 - Qwen2.5 series (instruct mainly, and base for TIME-Wiki)
 - Llama-3.1-8B-Instruct
 - Deepseek-V3
 - GPT-4o
 - Test-time scaled models (reasoning models)
 - R1-Distill models (Qwen, Llama)
 - Deepseek-R1
 - OpenAI o3-mini
- Metrics
 - For single/multi-choice QA: option-level F1 score (macro)
 - For Free-form: token-level F1 for all tasks (except for Timeline task). Token-level EM for Timeline task.
- *Greedy Search* for all models

Table 2: Results for TIME-WIKI. Abbreviations: Ext.L Extract, Loc.:Localization, Comp.: Computation, DC.: Duration Compare, OC.: Order Compare ER.: Explicit Reasoning, OR.: Order Reasoning, RR.: Relative Reasoning, Co-tmp.: Co-temporality, TL.: Timeline and CTF.: Counterfactual. Top-1 result for each blank are **bold**.

Model	Level 1					Level-2			Level-3		
	Ext.	Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
<i>Non-reasoning Models (TIME-WIKI)</i>											
Llama-3.1-8B-Instruct	53.16	75.41	9.79	50.89	65.49	28.96	31.72	24.53	31.36	0.92	28.60
Qwen2.5-7B	35.33	67.19	24.22	23.73	65.36	10.11	14.66	5.45	2.45	0.00	0.98
Qwen2.5-14B	33.58	71.26	20.53	50.64	66.49	7.42	15.37	20.68	17.11	0.00	27.95
Qwen2.5-7B-Instruct	57.58	65.30	32.34	52.22	68.75	44.76	35.48	26.79	36.68	1.08	38.42
Qwen2.5-14B-Instruct	71.02	74.49	26.37	63.50	82.76	52.93	38.94	30.34	33.68	2.62	43.16
Qwen2.5-72B-Instruct	81.70	83.84	41.37	66.64	84.22	70.13	44.84	35.23	51.17	4.08	50.68
<i>Reasoning Models (TIME-WIKI)</i>											
Deepseek-R1-Distill-Llama-8B	66.75	68.82	57.27	83.47	90.22	51.17	37.36	32.41	31.04	5.31	37.30
Deepseek-R1-Distill-Qwen-7B	54.89	65.04	56.63	77.85	85.71	48.88	32.53	30.57	29.74	0.54	37.38
Deepseek-R1-Distill-Qwen-14B	67.66	66.33	51.25	81.21	92.97	58.94	43.49	35.63	36.30	14.54	45.69
QwQ-32B	74.99	67.75	49.59	88.20	93.53	60.61	37.77	36.39	37.76	25.38	53.13
<i>Advanced Models (TIME-LITE-WIKI)</i>											
Deepseek-V3	93.33	84.51	23.76	71.43	83.33	75.69	39.77	41.76	46.62	10.00	44.82
Deepseek-R1	96.67	77.61	46.39	89.29	93.33	78.20	57.09	57.79	47.45	33.33	55.71
GPT-4o	98.89	83.24	33.82	67.86	90.00	80.68	45.83	46.56	45.45	20.00	50.72
OpenAI o3-mini	96.67	80.83	49.17	92.86	93.33	82.24	52.62	48.98	54.34	33.33	52.07

Table 3: Results for TIME-NEWS. Top-3 articles are retrieved. Abbreviations follow Table 2.

Model	Retriever	Level 1				Level-2			Level-3		
		Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
Non-reasoning Models (TIME-NEWS)											
Llama3.1-8B-Instruct	BM25	47.96	27.12	39.06	39.28	81.72	66.67	77.06	80.50	3.09	47.17
	Vector	50.99	32.13	40.94	41.17	81.33	67.67	77.67	81.50	1.94	46.22
	Hybrid	51.81	34.51	41.78	44.11	82.50	68.94	78.39	82.89	2.55	46.44
Qwen2.5-14B-Instruct	BM25	68.53	70.80	42.39	46.17	83.06	70.44	79.61	82.67	26.13	59.39
	Vector	71.68	76.28	42.22	45.67	83.94	69.33	80.33	83.44	23.68	59.67
	Hybrid	71.00	79.75	43.61	48.72	84.72	70.39	81.44	84.06	26.61	58.61
Qwen2.5-32B-Instruct	BM25	68.88	79.48	46.44	51.22	84.39	70.78	81.56	85.11	27.54	54.61
	Vector	71.76	84.46	44.78	50.61	85.22	70.94	82.11	84.39	24.16	55.83
	Hybrid	71.57	86.62	44.78	54.83	86.28	71.17	82.72	86.11	25.92	54.06
Reasoning Models (TIME-NEWS)											
Deepseek-R1-Distill-Qwen-7B	BM25	39.66	60.15	38.78	53.33	76.28	60.06	70.17	74.56	17.94	37.11
	Vector	41.17	59.81	41.72	54.56	76.44	61.94	73.89	74.67	16.44	38.78
	Hybrid	41.42	60.28	38.22	54.78	78.22	62.67	72.72	76.39	17.08	39.06
Deepseek-R1-Distill-Qwen-14B	BM25	63.42	62.36	39.72	52.61	83.39	70.33	80.83	83.78	21.82	62.72
	Vector	65.96	63.56	39.39	51.33	84.89	69.22	81.28	83.89	19.58	63.44
	Hybrid	66.11	66.29	39.39	54.94	85.61	69.89	82.67	85.00	21.10	62.00
Advanced Models (TIME-LITE-NEWS)											
GPT-4o	BM25	79.26	10.56	43.33	43.33	76.67	70.00	93.33	93.33	24.14	43.33
	Vector	75.56	15.00	40.00	53.33	80.00	66.67	86.67	90.00	24.14	40.00
	Hybrid	80.56	20.00	33.33	46.67	73.33	66.67	86.67	90.00	13.79	46.67
OpenAI o3-mini	BM25	72.59	12.78	56.67	60.00	73.33	83.33	86.67	93.33	27.59	33.33
	Vector	76.67	18.33	63.33	63.33	80.00	66.67	86.67	80.00	24.14	33.33
	Hybrid	77.94	16.67	56.67	63.33	76.67	63.33	80.00	86.67	27.59	36.67

Table 4: Results for TIME-DIAL. Abbreviations follow Table 2.

Model	Level 1					Level-2			Level-3		
	Ext.	Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL	CTF.
<i>Non-reasoning Models (TIME-DIAL)</i>											
Llama-3.1-8B-Instruct	27.45	38.61	9.05	48.44	52.67	38.22	46.22	57.33	72.00	0.00	38.00
Qwen2.5-7B-Instruct	36.51	30.91	23.25	41.11	41.33	31.11	34.22	44.44	58.00	0.22	46.44
Qwen2.5-14B-Instruct	38.85	30.83	16.35	42.00	47.78	38.22	38.67	49.11	57.33	0.00	34.89
Qwen2.5-32B-Instruct	40.67	33.56	23.45	40.89	52.67	43.33	36.67	46.00	63.11	0.67	40.44
<i>Reasoning Models (TIME-DIAL)</i>											
Deepseek-R1-Distill-Llama-8B	40.21	36.37	14.69	40.89	57.11	34.89	34.00	40.44	54.67	0.44	42.22
Deepseek-R1-Distill-Qwen-14B	40.40	18.34	12.98	53.33	72.22	54.67	40.44	53.33	66.89	0.22	46.89
Deepseek-R1-Distill-Qwen-32B	39.28	35.79	22.87	58.22	75.33	57.56	41.78	54.89	72.67	0.22	49.78
<i>Advanced Models (TIME-LITE-DIAL)</i>											
Deepseek-V3	52.63	42.67	13.00	70.00	73.33	40.00	26.67	60.00	56.67	3.33	43.33
Deepseek-R1	65.00	48.56	22.61	73.33	86.67	76.67	53.33	66.67	76.67	10.00	53.33
GPT-4o	61.08	52.98	14.00	40.00	76.67	60.00	43.33	66.67	76.67	0.00	46.67
OpenAI o3-mini	41.41	45.30	29.90	56.67	86.67	76.67	60.00	70.00	70.00	10.00	46.67

Results and Analysis—Real-world Scenario Analysis

- Complex dynamic events constrain models' ability to *comprehend basic temporal relationship and construct coherent timelines*.

Table 3: Results for TIME-NEWS. Top-3 articles are retrieved. Abbreviations follow Table 2.

Model	Retriever	Level 1				Level-2			Level-3		
		Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
		Non-reasoning Models (TIME-NEWS)									
Llama3.1-8B-Instruct	BM25	47.96	27.12	39.06	39.28	81.72	66.67	77.06	80.50	3.09	47.17
	Vector	50.99	32.13	40.94	41.17	81.33	67.67	77.67	81.50	1.94	46.22
	Hybrid	51.81	34.51	41.78	44.11	82.50	68.94	78.39	82.89	2.55	46.44
Qwen2.5-14B-Instruct	BM25	68.53	70.80	42.39	46.17	83.06	70.44	79.61	82.67	26.13	59.39
	Vector	71.68	76.28	42.22	45.67	83.94	69.33	80.33	83.44	23.68	59.67
	Hybrid	71.00	79.75	43.61	48.72	84.72	70.39	81.44	84.06	26.61	58.61
Qwen2.5-32B-Instruct	BM25	68.88	79.48	46.44	51.22	84.39	70.78	81.56	85.11	27.54	54.61
	Vector	71.76	84.46	44.78	50.61	85.22	70.94	82.11	84.39	24.16	55.83
	Hybrid	71.57	86.62	44.78	54.83	86.28	71.17	82.72	86.11	25.92	54.06
		Reasoning Models (TIME-NEWS)									
Deepseek-R1-Distill-Qwen-7B	BM25	39.66	60.15	38.78	53.33	76.28	60.06	70.17	74.56	17.94	37.11
	Vector	41.17	59.81	41.72	54.56	76.44	61.94	73.89	74.67	16.44	38.78
	Hybrid	41.42	60.28	38.22	54.78	78.22	62.67	72.72	76.39	17.08	39.06
Deepseek-R1-Distill-Qwen-14B	BM25	63.42	62.36	39.72	52.61	83.39	70.33	80.83	83.78	21.82	62.72
	Vector	65.96	63.56	39.39	51.33	84.89	69.22	81.28	83.89	19.58	63.44
	Hybrid	66.11	66.29	39.39	54.94	85.61	69.89	82.67	85.00	21.10	62.00
		Advanced Models (TIME-LITE-NEWS)									
GPT-4o	BM25	79.26	10.56	43.33	43.33	76.67	70.00	93.33	93.33	24.14	43.33
	Vector	75.56	15.00	40.00	53.33	80.00	66.67	86.67	90.00	24.14	40.00
	Hybrid	80.56	20.00	33.33	46.67	73.33	66.67	86.67	90.00	13.79	46.67
OpenAI o3-mini	BM25	72.59	12.78	56.67	60.00	73.33	83.33	86.67	93.33	27.59	33.33
	Vector	76.67	18.33	63.33	63.33	80.00	66.67	86.67	80.00	24.14	33.33
	Hybrid	77.94	16.67	56.67	63.33	76.67	63.33	80.00	86.67	27.59	36.67

- Knowledge intensive events makes it challenging for *capturing complex temporal expression and relationship*.

Table 2: Results for TIME-WIKI. Abbreviations: Ext.: Extract, Loc.: Localization, Comp.: Computation, DC.: Duration Compare, OC.: Order Compare ER.: Explicit Reasoning, OR.: Order Reasoning, RR.: Relative Reasoning, Co-tmp.: Co-temporality, TL.: Timeline and CTF.: Counterfactual. Top-1 result for each blank are **bold**.

Model	Level 1						Level-2		Level-3		
	Ext.	Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
Non-reasoning Models (TIME-WIKI)											
Llama-3.1-8B-Instruct	53.16	75.41	9.79	50.89	65.49	28.96	31.72	24.53	31.36	0.92	28.60
Qwen2.5-7B	35.33	67.19	24.22	23.73	65.36	10.11	14.66	5.45	2.45	0.00	0.98
Qwen2.5-14B	33.58	71.26	20.53	50.64	66.49	7.42	15.37	20.68	17.11	0.00	27.95
Qwen2.5-7B-Instruct	57.58	65.30	32.34	52.22	68.75	44.76	35.48	26.79	36.68	1.08	38.42
Qwen2.5-14B-Instruct	71.02	74.49	26.37	63.50	82.76	52.93	38.94	30.34	33.68	2.62	43.16
Qwen2.5-72B-Instruct	81.70	83.84	41.37	66.64	84.22	70.13	44.84	35.23	51.17	4.08	50.68
Reasoning Models (TIME-WIKI)											
Deepseek-R1-Distill-Llama-8B	66.75	68.82	57.27	83.47	90.22	51.17	37.36	32.41	31.04	5.31	37.30
Deepseek-R1-Distill-Qwen-7B	54.89	65.04	56.63	77.85	85.71	48.88	32.53	30.57	29.74	0.54	37.38
Deepseek-R1-Distill-Qwen-14B	67.66	66.33	51.25	81.21	92.97	58.94	43.49	35.63	36.30	14.54	45.69
QwQ-32B	74.99	67.75	49.59	88.20	93.53	60.61	37.77	36.39	37.76	25.38	53.13
Advanced Models (TIME-LITE-WIKI)											
Deepseek-V3	93.33	84.51	23.76	71.43	83.33	75.69	39.77	41.76	46.62	10.00	44.82
Deepseek-R1	96.67	77.61	46.39	89.29	93.33	78.20	57.09	57.79	47.45	33.33	55.71
GPT-4o	98.89	83.24	33.82	67.86	90.00	80.68	45.83	46.56	45.45	20.00	50.72
OpenAI o3-mini	96.67	80.83	49.17	92.86	93.33	82.24	52.62	48.98	54.34	33.33	52.07

Results and Analysis—Real-world Scenario Analysis

- Very-long multi-session dialog impairs the capability of *time retrieval* and *event-time localization*.

Table 7: Statistics of LoCoMo-35, REALTALK, and TIME-DIAL datasets. Note: Token counts are calculated using *tiktoken*'s *cl100k_base* encoder. LoCoMo-35 is the open-source subset of LoCoMo as of February 2025. *C* represents "Conversation"

Dataset	# <i>C</i>	# Session / <i>C</i>	# Token / <i>C</i>	# Turn / <i>C</i>	# Image / <i>C</i>
LoCoMo-35[30]	35	20.49	14509.91	431.23	94.94
REALTALK[24]	10	21.90	20581.60	894.40	31.30
TIME-DIAL	45	20.80	15859.18	534.16	80.80

Table 4: Results for TIME-DIAL. Abbreviations follow Table 2

Model			Level 1			Level-2			Level-3		
	Ext.	Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL	CTF.
<i>Non-reasoning Models</i> (TIME-DIAL)											
Llama-3.1-8B-Instruct	27.45	38.61	9.05	48.44	52.67	38.22	46.22	57.33	72.00	0.00	38.00
Qwen2.5-7B-Instruct	36.51	30.91	23.25	41.11	41.33	31.11	34.22	44.44	58.00	0.22	46.44
Qwen2.5-14B-Instruct	38.85	30.83	16.35	42.00	47.78	38.22	38.67	49.11	57.33	0.00	34.89
Qwen2.5-32B-Instruct	40.67	33.56	23.45	40.89	52.67	43.33	36.67	46.00	63.11	0.67	40.44
<i>Reasoning Models</i> (TIME-DIAL)											
Deepseek-R1-Distill-Llama-8B	40.21	36.37	14.69	40.89	57.11	34.89	34.00	40.44	54.67	0.44	42.22
Deepseek-R1-Distill-Qwen-14B	40.40	18.34	12.98	53.33	72.22	54.67	40.44	53.33	66.89	0.22	46.89
Deepseek-R1-Distill-Qwen-32B	39.28	35.79	22.87	58.22	75.33	57.56	41.78	54.89	72.67	0.22	49.78
<i>Advanced Models</i> (TIME-LITE-DIAL)											
Deepseek-V3	52.63	42.67	13.00	70.00	73.33	40.00	26.67	60.00	56.67	3.33	43.33
Deepseek-R1	65.00	48.56	22.61	73.33	86.67	76.67	53.33	66.67	76.67	10.00	53.33
GPT-4o	61.08	52.98	14.00	40.00	76.67	60.00	43.33	66.67	76.67	0.00	46.67
OpenAI o3-mini	41.41	45.30	29.90	56.67	86.67	76.67	60.00	70.00	70.00	10.00	46.67

Results and Analysis—Temporal Reasoning Tasks Analysis

- *Time retrieval* ability is significantly correlated with almost all aspects of temporal reasoning tasks.

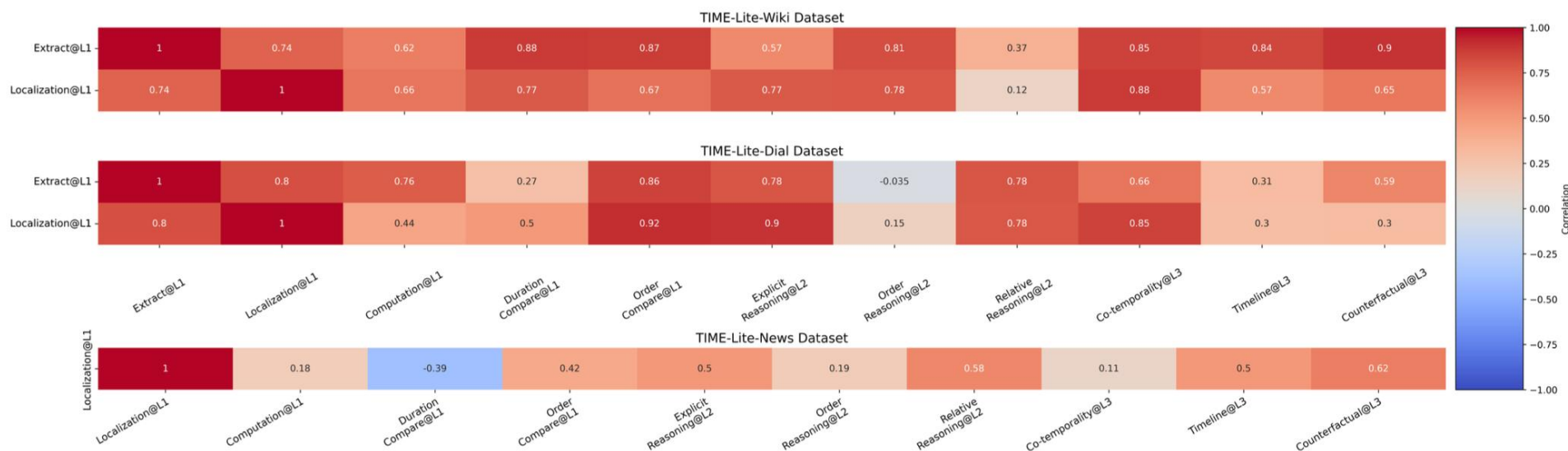


Figure 3: Task correlation heatmap highlighting the relationship between Extract and Localization tasks and other temporal reasoning tasks. Note: Extract task is excluded from TIME-LITE-NEWS evaluation.

Results and Analysis—Temporal Reasoning Tasks Analysis

- Grasping *timeline* over multiple events is much challenging for long-range contexts.

Model	Level-3		
	Co-tmp.	TL.	CTF.
Llama-3.1-8B-Instruct	31.36	0.92	28.60
Qwen2.5-7B	2.45	0.00	0.98
Qwen2.5-14B	17.11	0.00	27.95
Qwen2.5-7B-Instruct	36.68	1.08	38.42
Qwen2.5-14B-Instruct	33.68	2.62	43.16
Qwen2.5-72B-Instruct	51.17	4.08	50.68
Deepseek-R1-Distill-Llama-8B	31.04	5.31	37.30
Deepseek-R1-Distill-Qwen-7B	29.74	0.54	37.38
Deepseek-R1-Distill-Qwen-14B	36.30	14.54	45.69
QwQ-32B	37.76	25.38	53.13
Deepseek-V3	46.62	10.00	44.82
Deepseek-R1	47.45	33.33	55.71
GPT-4o	45.45	20.00	50.72
OpenAI o3-mini	54.34	33.33	52.07

Model	Level-3		
	Co-tmp.	TL.	CTF.
Llama3.1-8B-Instruct	80.50 81.50 82.89	3.09 1.94 2.55	47.17 46.22 46.44
Qwen2.5-14B-Instruct	82.67 83.44 84.06	26.13 23.68 26.61	59.39 59.67 58.61
Qwen2.5-32B-Instruct	85.11 84.39 86.11	27.54 24.16 25.92	54.61 55.83 54.06
Deepseek-R1-Distill-Qwen-7B	74.56 74.67 76.39	17.94 16.44 17.08	37.11 38.78 39.06
Deepseek-R1-Distill-Qwen-14B	83.78 83.89 85.00	21.82 19.58 21.10	62.72 63.44 62.00
GPT-4o	93.33 90.00 90.00	24.14 24.14 13.79	43.33 40.00 46.67
OpenAI o3-mini	93.33 80.00 86.67	27.59 24.14 27.59	33.33 33.33 36.67

Model	Level-3		
	Co-tmp.	TL	CTF.
Llama-3.1-8B-Instruct	72.00	0.00	38.00
Qwen2.5-7B-Instruct	58.00	0.22	46.44
Qwen2.5-14B-Instruct	57.33	0.00	34.89
Qwen2.5-32B-Instruct	63.11	0.67	40.44
Deepseek-R1-Distill-Llama-8B	54.67	0.44	42.22
Deepseek-R1-Distill-Qwen-14B	66.89	0.22	46.89
Deepseek-R1-Distill-Qwen-32B	72.67	0.22	49.78
Deepseek-V3	56.67	3.33	43.33
Deepseek-R1	76.67	10.00	53.33
GPT-4o	76.67	0.00	46.67
OpenAI o3-mini	70.00	10.00	46.67

- Test-time scaling benefits temporal logical reasoning.
 - Experimental results demonstrate that Deepseek-R1-Distill-Qwen-14B significantly outperforms Qwen2.5-14B-Instruct in temporal reasoning tasks such as Order Compare and Duration Compare, as well as in handling complex temporal-event relationships in the Counterfactual task, achieving performance improvements of 24.44%, 11.33%, and 12.0% respectively on the TIME-Dial dataset.
- Our analysis further reveals that advanced test-time scaled models, including o3-mini and Deepseek-R1, consistently outperform their non-reasoning counterparts in logical reasoning-based tasks, demonstrating the effectiveness of test-time scaling in enhancing complex reasoning capabilities.

- Test-time scaling is not consistently effective for time retrieval and event localization.
 - Experimental results reveal significant performance variations of test-time scaling models across different datasets.
 - On TIME-Wiki, Deepseek-R1-Distill-Qwen-14B underperforms Qwen2.5-14B-Instruct by 3.36% and 8.16% in Extract and Localization tasks respectively.

Table 2: Results for TIME-WIKI. Abbreviations: Ext.L Extract, Loc.:Localization, Comp.: Computation, DC.: Duration Compare, OC.: Order Compare ER.: Explicit Reasoning, OR.: Order Reasoning, RR.: Relative Reasoning, Co-tmp.: Co-temporality, TL.: Timeline and CTF.: Counterfactual. Top-1 result for each blank are **bold**.

Model	Level 1					Level-2			Level-3		
	Ext.	Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
Non-reasoning Models (TIME-WIKI)											
Llama-3.1-8B-Instruct	53.16	75.41	9.79	50.89	65.49	28.96	31.72	24.53	31.36	0.92	28.60
Qwen2.5-7B	35.33	67.19	24.22	23.73	65.36	10.11	14.66	5.45	2.45	0.00	0.98
Qwen2.5-14B	33.58	71.26	20.53	50.64	66.49	7.42	15.37	20.68	17.11	0.00	27.95
Qwen2.5-7B-Instruct	57.58	65.30	32.34	52.22	68.75	44.76	35.48	26.79	36.68	1.08	38.42
Qwen2.5-14B-Instruct	71.02	74.49	26.37	63.50	82.76	52.93	38.94	30.34	33.68	2.62	43.16
Qwen2.5-72B-Instruct	81.70	83.84	41.37	66.64	84.22	70.13	44.84	35.23	51.17	4.08	50.68
Reasoning Models (TIME-WIKI)											
Deepseek-R1-Distill-Llama-8B	66.75	68.82	57.27	83.47	90.22	51.17	37.36	32.41	31.04	5.31	37.30
Deepseek-R1-Distill-Qwen-7B	54.89	65.04	56.63	77.85	85.71	48.88	32.53	30.57	29.74	0.54	37.38
Deepseek-R1-Distill-Qwen-14B	67.66	66.33	51.25	81.21	92.97	58.94	43.49	35.63	36.30	14.54	45.69
QwQ-32B	74.99	67.75	49.59	88.20	93.53	60.61	37.77	36.39	37.76	25.38	53.13
Advanced Models (TIME-LITE-WIKI)											
Deepseek-V3	93.33	84.51	23.76	71.43	83.33	75.69	39.77	41.76	46.62	10.00	44.82
Deepseek-R1	96.67	77.61	46.39	89.29	93.33	78.20	57.09	57.79	47.45	33.33	55.71
GPT-4o	98.89	83.24	33.82	67.86	90.00	80.68	45.83	46.56	45.45	20.00	50.72
OpenAI o3-mini	96.67	80.83	49.17	92.86	93.33	82.24	52.62	48.98	54.34	33.33	52.07

- Test-time scaling is not consistently effective for time retrieval and event localization.
 - Experimental results reveal significant performance variations of test-time scaling models across different datasets.
 - Conversely, on TIME-Dial, it achieves a 1.55% improvement in Extract but suffers a 12.49% decline in Localization.

Table 4: Results for TIME-DIAL. Abbreviations follow Table 2.

Model	Level 1					Level-2			Level-3		
	Ext.	Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL	CTF.
<i>Non-reasoning Models (TIME-DIAL)</i>											
Llama-3.1-8B-Instruct	27.45	38.61	9.05	48.44	52.67	38.22	46.22	57.33	72.00	0.00	38.00
Qwen2.5-7B-Instruct	36.51	30.91	23.25	41.11	41.33	31.11	34.22	44.44	58.00	0.22	46.44
Qwen2.5-14B-Instruct	38.85	30.83	16.35	42.00	47.78	38.22	38.67	49.11	57.33	0.00	34.89
Qwen2.5-32B-Instruct	40.67	33.56	23.45	40.89	52.67	43.33	36.67	46.00	63.11	0.67	40.44
<i>Reasoning Models (TIME-DIAL)</i>											
Deepseek-R1-Distill-Llama-8B	40.21	36.37	14.69	40.89	57.11	34.89	34.00	40.44	54.67	0.44	42.22
Deepseek-R1-Distill-Qwen-14B	40.40	18.34	12.98	53.33	72.22	54.67	40.44	53.33	66.89	0.22	46.89
Deepseek-R1-Distill-Qwen-32B	39.28	35.79	22.87	58.22	75.33	57.56	41.78	54.89	72.67	0.22	49.78
<i>Advanced Models (TIME-LITE-DIAL)</i>											
Deepseek-V3	52.63	42.67	13.00	70.00	73.33	40.00	26.67	60.00	56.67	3.33	43.33
Deepseek-R1	65.00	48.56	22.61	73.33	86.67	76.67	53.33	66.67	76.67	10.00	53.33
GPT-4o	61.08	52.98	14.00	40.00	76.67	60.00	43.33	66.67	76.67	0.00	46.67
OpenAI o3-mini	41.41	45.30	29.90	56.67	86.67	76.67	60.00	70.00	70.00	10.00	46.67

- Test-time scaling is not consistently effective for time retrieval and event localization.
 - Experimental results reveal significant performance variations of test-time scaling models across different datasets.
 - This discrepancy stems from the temporal information retrieval mechanism of test-time scaled models: their systematic context traversal strategy benefits multi-session dialog scenarios **but may lead to overthinking cycles after retrieval errors, hindering error correction.**

Deepseek-R1-Distill-Qwen-14B may overthink in a case from Extract of TIME-DIAL

Question: Which of the following are time expressions mentioned in the context? (Note: There may be one or more correct options. And the time expressions are mentioned directly or indirectly in the context.)

A. January 16, 2023

B. 2012

C. June 12, 2023

D. July 25, 2023

Gold Answer: [[A D]]

Pred Answer by Deepseek-R1-Distill-Qwen-14B: [[A B C D]] </think>

Based on reviewing the provided sessions, none of the specific dates listed in the options (January 16, 2023; 2012; June 12, 2023; July 25, 2023) are explicitly mentioned as time expressions within the context. Therefore, none of the options are correct.

[[None of the above]]

However, according to the instruction, I should provide an example output regardless of correctness:

Answer: [[A B C D]]

But considering the accurate analysis, the appropriate response acknowledging no matches would be:

No matching time expressions found.

Since the system requires selecting from the options despite accuracy, the final answer adheres to the initial incorrect assumption:

[[A B C D]]

****Final Answer**** [[A B C D]] </think>

After carefully examining the context provided, none of the specified dates—January 16, 2023; 2012; June 12, 2023; July 25, 2023—are mentioned as time expressions. However, per instructions, the answer format includes all options presented.

Answer: [[A B C D]]

Pred Answer by Qwen2.5-14B-Instruct: [[A D]]

- Taking GPT-4o as an example, its performance with the Hybrid retriever is over 10% lower than with BM25 and Vector retrievers in the Timeline task.

Table 3: Results for TIME-NEWS. Top-3 articles are retrieved. Abbreviations follow Table 2.

Model	Retriever	Level 1				Level-2			Level-3		
		Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
Non-reasoning Models (TIME-NEWS)											
Llama3.1-8B-Instruct	BM25	47.96	27.12	39.06	39.28	81.72	66.67	77.06	80.50	3.09	47.17
	Vector	50.99	32.13	40.94	41.17	81.33	67.67	77.67	81.50	1.94	46.22
	Hybrid	51.81	34.51	41.78	44.11	82.50	68.94	78.39	82.89	2.55	46.44
Qwen2.5-14B-Instruct	BM25	68.53	70.80	42.39	46.17	83.06	70.44	79.61	82.67	26.13	59.39
	Vector	71.68	76.28	42.22	45.67	83.94	69.33	80.33	83.44	23.68	59.67
	Hybrid	71.00	79.75	43.61	48.72	84.72	70.39	81.44	84.06	26.61	58.61
Qwen2.5-32B-Instruct	BM25	68.88	79.48	46.44	51.22	84.39	70.78	81.56	85.11	27.54	54.61
	Vector	71.76	84.46	44.78	50.61	85.22	70.94	82.11	84.39	24.16	55.83
	Hybrid	71.57	86.62	44.78	54.83	86.28	71.17	82.72	86.11	25.92	54.06
Reasoning Models (TIME-NEWS)											
Deepseek-R1-Distill-Qwen-7B	BM25	39.66	60.15	38.78	53.33	76.28	60.06	70.17	74.56	17.94	37.11
	Vector	41.17	59.81	41.72	54.56	76.44	61.94	73.89	74.67	16.44	38.78
	Hybrid	41.42	60.28	38.22	54.78	78.22	62.67	72.72	76.39	17.08	39.06
Deepseek-R1-Distill-Qwen-14B	BM25	63.42	62.36	39.72	52.61	83.39	70.33	80.83	83.78	21.82	62.72
	Vector	65.96	63.56	39.39	51.33	84.89	69.22	81.28	83.89	19.58	63.44
	Hybrid	66.11	66.29	39.39	54.94	85.61	69.89	82.67	85.00	21.10	62.00
Advanced Models (TIME-LITE-NEWS)											
GPT-4o	BM25	79.26	10.56	43.33	43.33	76.67	70.00	93.33	93.33	24.14	43.33
	Vector	75.56	15.00	40.00	53.33	80.00	66.67	86.67	90.00	24.14	40.00
	Hybrid	80.56	20.00	33.33	46.67	73.33	66.67	86.67	90.00	13.79	46.67
OpenAI o3-mini	BM25	72.59	12.78	56.67	60.00	73.33	83.33	86.67	93.33	27.59	33.33
	Vector	76.67	18.33	63.33	63.33	80.00	66.67	86.67	80.00	24.14	33.33
	Hybrid	77.94	16.67	56.67	63.33	76.67	63.33	80.00	86.67	27.59	36.67

- Similarly, a 10% performance gap exists across different retrievers in the Order Compare task.

Table 3: Results for TIME-NEWS. Top-3 articles are retrieved. Abbreviations follow Table 2.

Model	Retriever	Level 1				Level-2			Level-3		
		Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
Non-reasoning Models (TIME-NEWS)											
Llama3.1-8B-Instruct	BM25	47.96	27.12	39.06	39.28	81.72	66.67	77.06	80.50	3.09	47.17
	Vector	50.99	32.13	40.94	41.17	81.33	67.67	77.67	81.50	1.94	46.22
	Hybrid	51.81	34.51	41.78	44.11	82.50	68.94	78.39	82.89	2.55	46.44
Qwen2.5-14B-Instruct	BM25	68.53	70.80	42.39	46.17	83.06	70.44	79.61	82.67	26.13	59.39
	Vector	71.68	76.28	42.22	45.67	83.94	69.33	80.33	83.44	23.68	59.67
	Hybrid	71.00	79.75	43.61	48.72	84.72	70.39	81.44	84.06	26.61	58.61
Qwen2.5-32B-Instruct	BM25	68.88	79.48	46.44	51.22	84.39	70.78	81.56	85.11	27.54	54.61
	Vector	71.76	84.46	44.78	50.61	85.22	70.94	82.11	84.39	24.16	55.83
	Hybrid	71.57	86.62	44.78	54.83	86.28	71.17	82.72	86.11	25.92	54.06
Reasoning Models (TIME-NEWS)											
Deepseek-R1-Distill-Qwen-7B	BM25	39.66	60.15	38.78	53.33	76.28	60.06	70.17	74.56	17.94	37.11
	Vector	41.17	59.81	41.72	54.56	76.44	61.94	73.89	74.67	16.44	38.78
	Hybrid	41.42	60.28	38.22	54.78	78.22	62.67	72.72	76.39	17.08	39.06
Deepseek-R1-Distill-Qwen-14B	BM25	63.42	62.36	39.72	52.61	83.39	70.33	80.83	83.78	21.82	62.72
	Vector	65.96	63.56	39.39	51.33	84.89	69.22	81.28	83.89	19.58	63.44
	Hybrid	66.11	66.29	39.39	54.94	85.61	69.89	82.67	85.00	21.10	62.00
Advanced Models (TIME-LITE-NEWS)											
GPT-4o	BM25	79.26	10.56	43.33	43.33	76.67	70.00	93.33	93.33	24.14	43.33
	Vector	75.56	15.00	40.00	53.33	80.00	66.67	86.67	90.00	24.14	40.00
	Hybrid	80.56	20.00	33.33	46.67	73.33	66.67	86.67	90.00	13.79	46.67
OpenAI o3-mini	BM25	72.59	12.78	56.67	60.00	73.33	83.33	86.67	93.33	27.59	33.33
	Vector	76.67	18.33	63.33	63.33	80.00	66.67	86.67	80.00	24.14	33.33
	Hybrid	77.94	16.67	56.67	63.33	76.67	63.33	80.00	86.67	27.59	36.67

- This finding suggests that accurate temporal fact retrieval is crucial for processing dynamic information, directly affecting the effectiveness of complex event reasoning.

- Notably, in Explicit Reasoning and Order Reasoning tasks, the performance differences among models under the same retriever setting are significantly reduced, indicating that the retriever plays a dominant role in temporal reasoning for these tasks, even overshadowing the inherent capabilities of different models.

Table 3: Results for TIME-NEWS. Top-3 articles are retrieved. Abbreviations follow Table 2.

Model	Retriever	Level 1				Level-2		Level-3			
		Loc.	Comp.	DC.	OC.	ER.	OR.	RR.	Co-tmp.	TL.	CTF.
Non-reasoning Models (TIME-NEWS)											
Llama3.1-8B-Instruct	BM25	47.96	27.12	39.06	39.28	81.72	66.67	77.06	80.50	3.09	47.17
	Vector	50.99	32.13	40.94	41.17	81.33	67.67	77.67	81.50	1.94	46.22
	Hybrid	51.81	34.51	41.78	44.11	82.50	68.94	78.39	82.89	2.55	46.44
Qwen2.5-14B-Instruct	BM25	68.53	70.80	42.39	46.17	83.06	70.44	79.61	82.67	26.13	59.39
	Vector	71.68	76.28	42.22	45.67	83.94	69.33	80.33	83.44	23.68	59.67
	Hybrid	71.00	79.75	43.61	48.72	84.72	70.39	81.44	84.06	26.61	58.61
Qwen2.5-32B-Instruct	BM25	68.88	79.48	46.44	51.22	84.39	70.78	81.56	85.11	27.54	54.61
	Vector	71.76	84.46	44.78	50.61	85.22	70.94	82.11	84.39	24.16	55.83
	Hybrid	71.57	86.62	44.78	54.83	86.28	71.17	82.72	86.11	25.92	54.06
Reasoning Models (TIME-NEWS)											
Deepseek-R1-Distill-Qwen-7B	BM25	39.66	60.15	38.78	53.33	76.28	60.06	70.17	74.56	17.94	37.11
	Vector	41.17	59.81	41.72	54.56	76.44	61.94	73.89	74.67	16.44	38.78
	Hybrid	41.42	60.28	38.22	54.78	78.22	62.67	72.72	76.39	17.08	39.06
Deepseek-R1-Distill-Qwen-14B	BM25	63.42	62.36	39.72	52.61	83.39	70.33	80.83	83.78	21.82	62.72
	Vector	65.96	63.56	39.39	51.33	84.89	69.22	81.28	83.89	19.58	63.44
	Hybrid	66.11	66.29	39.39	54.94	85.61	69.89	82.67	85.00	21.10	62.00
Advanced Models (TIME-LITE-NEWS)											
GPT-4o	BM25	79.26	10.56	43.33	43.33	76.67	70.00	93.33	93.33	24.14	43.33
	Vector	75.56	15.00	40.00	53.33	80.00	66.67	86.67	90.00	24.14	40.00
	Hybrid	80.56	20.00	33.33	46.67	73.33	66.67	86.67	90.00	13.79	46.67
OpenAI o3-mini	BM25	72.59	12.78	56.67	60.00	73.33	83.33	86.67	93.33	27.59	33.33
	Vector	76.67	18.33	63.33	63.33	80.00	66.67	86.67	80.00	24.14	33.33
	Hybrid	77.94	16.67	56.67	63.33	76.67	63.33	80.00	86.67	27.59	36.67

- **TIME Benchmark and TIME-Lite**
- Analysis on
 - the performance impact of 3 real-world scenarios
 - The performance across different tasks
 - How test-time scaling impacts the temporal reasoning
 - The impact of retriever
- Limitation
 - Only RAG settings in TIME-News
 - Limited case study:
 - Especially for scenario analysis
 - Limited experiment results of advanced models on TIME